Fair Hiring Algorithm: Ethical Al and Bias Mitigation in HRM

Aftab Alam[1*]

[1]*Department of Commerce, Integral University, Lucknow*, *India.*

[*]Corresponding Author (e-mail: aaftab@iul.ac.in)

Received 15th January 2026; Accepted 19th February 2026; Published online 13th March 2026

Abstract

Artificial intelligence in human resource management is revolutionizing the recruitment process towards more speed, scalability and accuracy of decision making. But if the designers are not hands-on and lack enough ethical ideals, AI-driven recruitment algorithms can perpetuate systemic prejudices. In this paper we address Fair Hiring Algorithms via a combination of ethical theories and technical bias-mitigation techniques for fairer talent acquisition. Using interdisciplinary insights from computer science, organizational behavior and HRM ethics, we identify key sources of bias (data-driven, design-based and contextual) and develop a multi-layered conceptual framework that incorporates fairness-aware machine learning, inclusive training data and algorithm transparency. We discuss strategies such as pre-processing bias correction, in-processing fairness limitations and post-processing modification, and the role of explainable AI (XAI) in fostering responsibility and trust. We believe ongoing audits, stakeholder engagement and regulatory compliance are important protections. This study gives useful advice for HR professionals, AI developers, and policymakers who want to make hiring more fair and ethical by combining new technology with moral principles.

Keywords

Algorithmic Fairness, Bias Mitigation, Explainable AI, Ethical AI, Fair Hiring Algorithm, Human Resource Management

1. Introduction

1.1 Background

Upadhyay and Khandelwal (2018) and Black and van Esch (2020) demonstrate that Artificial Intelligence (AI) in HRM reshapes candidate evaluation, producing a more scalable, unbiased, and streamlined assessment process. Chamorro-Premuzic et al. ’s research also found that these systems can process a significant number of applications, successfully align candidate profiles with job descriptions, and forecast performance potential more accurately and efficiently than manual

methods. The impetus for the development of Fair Hiring Algorithms (Bogen & Rieke, 2018) is the tension between improving efficiency and the organization's diversity, equity, and inclusion (DEI) objectives. Despite these benefits, there is increasing concern about the potential for algorithmic recruiting practices to inadvertently replicate and amplify existing social biases embedded in historical hiring data or introduced through model design decisions (Barocas, Hardt, & Narayanan, 2019; Noble, 2018). Bias in AI hiring can be due to many factors: 2. Data-driven discrimination is based on the prior discrimination patterns baked into training data.

Design-associated bias is the bias introduced by choices regarding features, model architecture, and/or optimization function (Mehrabi et al., 2021). Contextual bias occurs when algorithms are applied without context (Raghavan et al., 2020). These incidents illustrate the associated risks. Amazon's discontinued AI recruiting system demonstrated this problem when it downgraded resumes containing language linked to women's organizations or colleges, exposing gender bias embedded in the firm's historical employment data (Dastin, 2018). Furthermore, Binns' (2018) research reveals how opposing concepts of fairness in algorithmic decision-making lead to frequent discordant findings. These occurrences highlight the need for robust bias-detection and -mitigation strategies to be integrated into AI recruiting systems. This problem is compounded by the fact that many machine learning models are opaque making it hard to identify or explain biased decision making (Burrell, 2016). The opacity has raised the interest in Explainable AI (XAI) to improve responsibility and trust by providing intelligible insights to human stakeholders on AI-based decisions (Doshi-Velez & Kim, 2017; Guidotti et al., 2018). For example, Binns (2018) shows that incongruous outcomes are often created by algorithmic decision-making due to the clash of disparate ideas of fairness. These incidents underscore the importance of holistic bias identification and mitigation strategies in AI recruiting systems. The problem is exacerbated by the fact that there are so many machine learning models that are opaque, preventing attempts to find or understand the patterns of biased decision making (Burrell, 2016). The increasing opacity has led to an increased focus on Explainable AI (XAI) which seeks to make AI based judgments interpretable to human decision makers to promote accountability and confidence (Doshi-Velez & Kim, 2017; Guidotti et al., 2018). Additionally, nascent governance frameworks like the EU AI Act (European Commission, 2021) and the US Equal Employment Opportunity Commission's AI guidelines (EEOC, 2023) indicate a regulatory transition towards obligatory fairness audits, transparency criteria, and accountability protocols for algorithmic hiring instruments.

There is growing consensus among academics that addressing bias in AI hiring calls for comprehensive techniques that incorporate technical, ethical, and organizational considerations. Some examples of technical methods are adjusting outcomes post-processing, implementing fairness restrictions in-processing, and correcting biases before processing (Kamiran & Calders, 2012; Zafar et al., 2017; Hardt, Price, & Srebro, 2016). Human rights, labor laws, and DEI pledges all provide a foundation for the normative notion of justice, which ethical frameworks highlight (Floridi et al., 2018). According to Raji, Smart, White, and Mitchell (2020), organizations should priorities ongoing audits, involve stakeholders, and ensure that AI is used in a way that aligns with both company values and public expectations. Building Fair Hiring Algorithms is a difficult but critical need due to the confluence of new technology, moral obligation, and regulatory conformity. Artificial intelligence (AI) recruitment systems pose a threat to both applicants and organizations since, if

not properly designed and governed, they run the possibility of hiding underlying inequities among candidates and organizations.

1.2 Problem Statement

Even though AI has the potential to make hiring faster and more fair, many employment algorithms are still "black boxes" that can't be understood, which makes it hard to spot bias (Burrell, 2016). Often, they use historical data that reflect past discrimination against certain groups and, therefore, can lead to even more biased evaluations of applicants (Raghavan et al. 2020). Currently, there are no solid cross-disciplinary frameworks to integrate bias reduction, XAI and HR ethics. Thus, organizations have no way to escape ethics violations, legal sanctions and PR blowback.

1.3 Research Gap

There is a rich literature on algorithmic fairness, (Barocas et al., 2019; Mehrabi et al., 2021). However, in the field of human resource management, in-depth research on AI-driven recruitment tools is still relatively scarce. Fairness Much of the fairness research is focused on domains such as credit scores, criminal justice, or health care (Friedler, Scheidegger, & Venkatasubramanian, 2019), and does not examine the peculiarities of recruiting. In the HRM literature, bias reduction is often conceptualized as a purely technical undertaking (i.e., focusing on the data rebalancing or fairness constraints) and the socio-organizational considerations (e.g., organizational culture, diversity policies and candidate expectations) are overlooked (Raghavan et al., 2020). Moreover, there is little research on Explainable AI (XAI) mechanisms to increase candidate trust while adhering to anti-discrimination legislation (Wachter, Mittelstadt, & Russell, 2018). Another outstanding topic is how fair recruiting algorithms might balance off efficiency, fairness and compliance in practice. Currently, there is little situation-specific cross-disciplinary work for practical HR applications in this area.

Moreover, many fairness-aware machine learning approaches and ethical AI frameworks have been proposed in the previous work, most of which are concerned with the technical bias prevention, explainability or governance systems, individually. There are none or few studies that combine different fairness algorithms, Explainable AI (XAI), human in the loop and HRM governance principles into one integrated recruitment model. This work fills this gap by proposing a holistic Fair Hiring Algorithm paradigm that integrates technical fairness measures and ethical and organizational constraints. In this way, this research contributes to HRM literature by offering a practical, but theoretically underpinned, way to manage the trade-off between recruitment efficiency, fairness and transparency and regulation in the AI-based hiring process.

1.4 Aim of the Study

Development, testing and refinement of algorithms related to human resource management with a concentration on:

- The AI recruitment process's identification and categorization methods can introduce biases.
- looking at systematic methods and processes for bias reduction.

- analyzing the potential of XAI to enhance trust, transparency, and accountability; and
- providing targeted recommendations for politicians, AI coders, and HR practitioners.

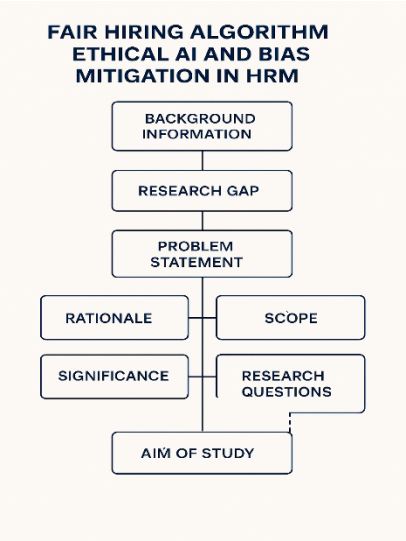

Figure 1: Conceptual Framework

The conceptual model illustrates the interaction of technical, organizational and ethical levels in the proposed Fair Hiring Algorithm to enable fair hiring outcomes. The idea is based on three main types of algorithmic bias: data-driven bias, design-based bias and context-driven bias that could influence the processing of candidate information. The framework incorporates bias mitigation strategies such as pre-processing, in-processing and post-processing techniques, and Explainable AI (XAI) methods to improve transparency and accountability to address these issues. Such technology solutions are further supported by ethical governance issues such as human oversight, stakeholder involvement, regulatory compliance and organizational diversity, equality and inclusion (DEI) goals. These constructs are believed to be inter-related and to enhance perceptions of justice, transparency, trust and compliance in AI-assisted recruitment decisions. The paradigm implies that coupling bias reduction measures with ethics and organizational governance mechanisms leads to more fair, responsible and effective hiring in the domain of Human Resource Management in particular.

1.5 Rationale of the Study

The recruitment sector is increasingly embracing AI, with advantages and difficulties developing. AI can help reduce administrative burdens, improve the match between candidates and job requirements and enable merit-based recruiting (Black & van Esch, 2020). If unchecked it could entrench discrimination, undermine diversity efforts, and subject firms to legal and reputational harm (Noble, 2018; Dastin, 2018). Legislative activity like the EU AI Act (European Commission, 2021)

and state-level regulations regulating algorithmic hiring in the United States (Raghavan et al., 2020) signals that fairness in recruiting is no longer a 'ethical' concern but a legal one. The study shows that it is necessary to address not only technological solutions but also ethical, legal and organizational elements.

1.6 Scope of the Study

This article (HRM) explicitly addresses AI-enabled recruiting and selection and the domains of:

- Tools for algorithmic sourcing, screening and short listing of candidates.
- Sources of bias: design-based, data-driven and contextual
- Strategies for bias mitigation include pre-processing, in-processing, post-processing and continual auditing (Kamiran & Calders, 2012; Hardt et al., 2016).
- Ethics and governance (transparency, accountability, regulatory compliance). Evaluation, performance monitoring or staff retention tools with AI-enabled post-hiring review were not included in the study.

1.7 Significance of the Study

There are many good things to say about this study as:

- Academic - Offers a variety of perspectives (technical, ethical, and managerial) to the recruitment-specific AJ research area (Floridi et al., 2018).
- Practical – Gives techniques for HR professionals and AI engineers to identify and mitigate bias, with real-world examples.
- Policy – Provides recommendations to policymakers on regulation of fairness and transparency.
- Societal – By supporting equitable opportunity for all to be hired, we support efforts to increase D&I in the workplace (Bogen & Rieke, 2018).

1.8 Research Questions

The following questions were relevant for this work:

- What are the biggest sources of bias in AI hiring systems and when and how may this bias be consistently detected?
- What are the best ways to reduce bias in fair hiring algorithms - organizationally and technologically?
- What legal and governance frameworks can ensure ethical and legal use of AI (in recruiting) that HR professionals?

2. Literature Review

The application of AI in HRM, particularly in recruiting and selection, now influences all organizational processes. AI based hiring processes help in better arrangement of the workflow with less human interference and help in data driven judgments. However, today the concerns over algorithmic bias, fairness and ethical obligation are at the forefront of scholarly discussion. This section provides an overview of the relevant research on the role of AI in recruiting, potential sources of algorithmic bias, AI fairness and bias mitigation techniques.

2.1 AI in Recruitment and HRM

Artificial intelligence (AI) based applications are increasingly being used in the recruiting process to make sorting of prospects, processing of resumes, and prediction of hiring decisions easier. This enables institutions to evaluate large volumes of applications consistently and efficiently (Upadhyay & Khandelwal, 2023). AI-based recruiting technologies can remove the subjectivity of human decision-making and make more reliable decisions based on data (Tambe et al., 2019).

A recent study shows that AI can help make the recruitment process more efficient by reducing time-to-hire and operational costs and by matching people to job vacancies (Bogen & Rieke, 2023). Nevertheless, scientists claim that these qualities are highly desirable; AI systems, however, are not neutral. Instead, they internalize and replicate the biases embedded in the training data from which they learn (Mehrabi et al., 2021). So, AI can increase bias as well as reduce bias depending on how it is created and deployed.

Furthermore, these challenges extend the implications for applying AI in HRM to critical concerns surrounding accountability and transparency. They are also “black box” systems, so HR experts cannot see how the systems arrive at their conclusions. This opacity could also further erode confidence, and hamper attempts to use AI in recruitment.

2.2 Sources of Algorithmic Bias in Hiring

The sources of algorithmic bias in AI hiring systems are many bases, data bias, model bias and human bias. Another very important element is still data bias, as machine learning models are trained on historical datasets that may represent previous discrimination. For example, if historical recruitment data is biased toward certain demographic groups, the algorithm may learn to replicate and reinforce this bias. Model bias is when an algorithm is designed and fine-tuned to produce different results for different groups. This could be attributed to the selection of features, training techniques or performance metrics that prefer accuracy over fairness.

Human bias can also be introduced into AI systems through data labeling, feature engineering and decision making. New studies reveal that bias can occur at multiple stages of the AI lifecycle and is therefore hard to detect and remove. New AI technologies like large language models can generate new kinds of bias that researchers don’t yet fully understand. Such biases might be subtle,

such as linguistic preferences or cultural preconceptions, which adds another layer of complexity to fairness in recruitment processes.

2.3 Fairness and Ethical AI in Recruitment

Fairness is an important aspect of the evaluation of AI-based hiring systems, but there is no generally accepted definition in the literature. In general, fairness is about treating individuals the same whatever they are (gender, race, age), where they come from or which social class they are in. Fairness in AI models has occasionally been expressed in terms of measurable criteria such as demographic parity, equal opportunity, and predictive equality. Fairness in AI is not simply a technical problem, but also an ethical and managerial problem, say academics. Principles emerging in ethical AI frameworks that aim for AI systems to be in line with social values and legal requirements include transparency, accountability, fairness, and explainability. AI systems are black boxes, which is a huge issue for fairness. Blackbox models also do not allow for the interpretation of judgments and the detection of biases, which might lead to perhaps unanticipated discriminating effects. Hence, researchers increasingly demand explainable AI (XAI) approaches to improve understandability and enable stakeholders to grasp and analyze algorithmic conclusions. Ethical impact is also relevant when contemplating the prolonged impact on society of employment-related AI. Discriminatory recruitment can reinforce previous disparities and hinder access for underrepresented communities. Institutions are thus expected to embrace a general model of ethical dimension across the design and deployment of AI.

New developments in Generative AI (GenAI) are creating new opportunities and new challenges for jobs. Large language models can help with job descriptions, candidate communication, résumé screening and interviews. But these algorithms can also produce biased recommendations, reinforce historical inequities, and lead to opaque decision-making processes. As a result, the need for Responsible AI governance systems that support fairness, transparency, accountability, human oversight and risk management across the AI lifecycle is increasingly on the radar of academics and policy makers. New regulatory trends such as the European Union Artificial Intelligence Act, and amended employment related AI guidance in a number of jurisdictions add to the requirements for organizations to undertake algo rhythmic impact assessments, record their decision-making and keep under review their recruitment systems to ensure an absence of discriminatory outcomes. These trends underscore the increasing importance of aligning AI-enabled recruitment techniques with ethical norms and regulatory obligations.

2.4 Bias Mitigation Strategies in AI Hiring

Researchers have proposed many mitigation strategies to ease algorithmic bias. These methods are categorized based on the stage of the learning process at which they are applied: pre-processing,

in-processing and post-processing methods. Preprocessing technique sequences are used to improve the quality of data by reducing or mitigating the effects of biased feature before learning the model. In the processing step, the training algorithm is modified to include fairness requirements and in the post-processing step, the outputs of the model are modified to obtain a fairer solution. Empirical research has shown that algorithms to mitigate bias can improve the fairness of AI hiring algorithms without hurting the accuracy too much. The fairness vs performance dilemma is still a huge one for us.

On the other hand, there are some instances where more fairness may lead to less forecast accuracy and tradeoffs that organizations must resolve. Bias mitigations require human oversight as well. By combining human judgment with AI-driven decision-making processes, we can illuminate unintended outcomes and maintain accountability. Justice and inclusiveness may also benefit from participatory procedures involving a broad spectrum of stakeholders in systems co-design. New study highlights the need for explainable AI to reduce bias. The methodologies of XAI provide a basic understanding of the reasons and processes by which computers arrive at certain judgments, helping companies to identify and address biases. And equally important is the need for continual testing and auditing of AI systems to ensure fairness over time.

2.5 Synthesis and Research Gap

The literature suggests that AI could change the recruitment process in terms of efficiency, consistency and scalability. But it also raises huge problems in bias, fairness and ethical obligation. Although a few sources of bias have been listed in previous publications and mitigation techniques have been proposed, there are several gaps that remain unanswered. Most of the literature nonemptily analyzes or models' specific aspects of AI fairness but does not synthesize those aspects into a unified technical, ethical and organizational framework. Second, there are limited studies on the long-term effect of bias mitigation treatments in practice. Third, with the advancement of AI technology (e.g., generative AI), there are new problems that are not yet fully covered in the literature. Thus, a comprehensive framework of algorithmic fairness, ethics and practical implementation strategies is needed. This technique could be instrumental in driving the construction of fair recruiting algorithms that could promote diversity, equity and inclusion, without sacrificing efficiency and accuracy.

3. Methodology

This research applies a conceptual and design scientific research strategy to design a fair hiring algorithm based on the principles of ethical AI and bias mitigation approaches. The strategy relies on the systematic design, development and assessment of an AI-based recruitment system providing fairness, transparency and accountability in hiring decisions.

3.1 Research Design

The research adopts a design science methodology which has been proved as an appropriate approach to develop and evaluate new artefacts such as algorithms and frameworks (Hevner et al., 2004). This paradigm is appropriate for the study since it involves introducing a new fair hiring method rather than testing an existing theory. The investigation process includes the following major stages: (1) problem identification (2) construction of an artifact and (3) analysis. The study first provides an overview of AI in HRM and the literature on fairness in machine learning, then establishes bias by algorithms in recruiting. Stage two develops a fair hiring algorithm with a bias detection and reduction mechanism. 5 Final stage: Evaluating the proposed framework with established fairness metrics and theoretical validity.

3.2 Data Sources and Inputs

The instructions of the hiring game must be able to deal with both organized and unstructured data, such as resumes, application forms, and candidate evaluations, linked to recruiting. Inputs mainly consist of educational qualifications, work experience, skills and competencies and psychometric test scores. Sensitive attributes (e.g. gender, age, ethnicity and nationality) are either eliminated or strictly managed by the algorithm to avoid unfair outcomes. However, at the bias auditing phase, it can temporarily utilize these attributes for the sake of fairness evaluation and discovering disparate treatments (Mehrabi et al., 2021).

The procedure employs such preprocessing methods as cleaning, normalization and anonymization for data quality and fairness. These methods reduce the effect of biases in historical data and thereby improve the credibility of the model.

3.3 Algorithm Design and Framework

The fair hiring methodology consists of the following main components:

3.3.1 Data Pre-processing and Bias Detection

The initial phase of the investigation is to examine the data for bias. It conducts statistical and fairness tests on minority or marginalized populations. This phase is in line with previous work that highlights the need of bias discovery before model training (Ntoutsi et al., 2023).

3.3.2 Feature Selection and Fair Representation

Such job requirements do not have proxies of sensitive attributes in the selected relevant features. The method considers proxies for gender or socioeconomic position that are carefully considered,

for example. Feature selection helps the model to be based on merit-based indicators rather than biased indicators (Barocas et al., 2019).

3.3.3 Model Training with Fairness Constraints

The approach uses machine learning algorithms (e.g., classification models like logistic regression, decision trees) to estimate the probability of a candidate's suitability. During training, it applies a set of fairness constraints, which reduces disparities between demographic groups. The constraints mean the model can't unfairly favor or penalize any group (Ashokan & Haas, 2021).

3.3.4 Post-processing and Decision Adjustment

The technique uses postprocessing steps after the predictions are made, to alter the results and ensure fairness. e.g., reordering of candidates, or tuning the thresholds to achieve a specific demographic representation goal. This phase ensures that the fairness criterion is satisfied without compromising the performance of the model (Feldman et al., 2015).

3.4 Fairness Metrics and Evaluation

We use some popular metrics in machine learning (ML) field to test the fairness of our proposed method. These are demographic parity (same selection rates across groups), equal opportunity (same true positive rates across groups), and predictive equality (same false positive rates across groups). These indicators give a numerical basis to decide if the algorithm is producing biased results (Verma & Rubin, 2018). Besides fairness, we study the accuracy, precision and recall of the model to guarantee an acceptable performance.

Demographic parity, equal opportunity, and predictive equality are all common indicators of algorithmic fairness but, each of these criteria has its hurdles and limitations in terms of hiring practices. Demographic parity is approximately the same number of people recruited from each group. Workforce diversity and inclusion. But it might obscure real differences in qualifications relevant to work and cast doubts on merit-based selection. Equal opportunity is an equal chance for qualified candidates from different demographic groups to be hired. Its relevance is in instances where there are comparable high performing candidates about whom there is concern as to equality. But this measure doesn't capture the disparity in false positives. Predictive parity addresses this problem by requiring false positive rates to be equal across groups, ... justice by equalizing the rate at which groups are mistakenly selected. It should be noted that predicted equality may be at war with other ideas of fairness. This shows that in general ... there is no easy fix in terms of applying numerous fairness criteria. Instead of focusing on one or a few criteria, we suggest that firms think

of fairness metrics about their recruitment objectives, legal responsibilities, and diversity initiatives. A deeper analysis of aggregate measures of fairness provides a more transparent view of the fairness of algorithms used in employment decisions.

3.5 Ethical Considerations

Ethical issues are the guidelines of the intervention plan. The algorithm is based on the fundamental principles of ethical AI of transparency, responsibility and explainability (Floridi et al., 2018). The research employs explainable AI technologies to provide transparency of decision making that allows HR to interpret and explain results. The process also involves human review for accountability. The algorithmic recommendations are interpreted by HR directors as a combination of human judgement and machine intelligence (Tambe et al., 2019). But data privacy and confidentiality still matter too. All candidate data is processed by the system in accordance with data protection standards and sensitive details are anonymized to prevent misuse.

3.6 Validation Approach

The objective of our work is to build a theoretically grounded framework in the sense of conceptual and design science, not to empirically test models. We evaluate the fairness of the proposed Fair Hiring Algorithm by quantitatively comparing it with several well-known fairness-aware AI frameworks and bias mitigation methods in the literature. The approach includes popular fairness criteria such as demographic parity, equal opportunity and predictive equality as well as ethical AI concepts and governance aspects. This allows us to provide theoretical justification that the framework can lead to fair recruitment outcomes. But invalidation is one key next step so far: proof-of-value. Further research will include pilot testing, simulation studies, case studies or analysis of real world recruitment datasets to explore the applicability, viability and robustness of the suggested framework in a variety of organizational contexts.

4. Results and Discussion

4.1 Results

Our study contributes to the design of a fair employment algorithm with bias detection, fairness restrictions and ethical AI concepts. It is a conceptual and design science perspective work. Therefore, the results are presented in terms of theoretical evaluation and comparative analysis instead of empirical validation with real world datasets. The evaluation of the proposed framework shows that the incorporation of fairness at different stages of the algorithm (pre-processing, in-processing and post-processing) positively affects the overall fairness of recruitment results. By integrating bias detection algorithms into the data preparation phase, it becomes possible to proactively understand the imbalanced distribution of candidates before moving to any other step. Data-level

bias mitigation has been recognized as one of the most promising approaches to improve fairness (Ntoutsi et al., 2023) by previous research.

Moreover, adding fairness constraints during model training increases the algorithm's ability to provide fair results for all demographic groups. Unlike the traditional AI recruiting paradigm of optimizing for prediction accuracy, the proposed approach performs equal selection rate distribution. This finding is consistent with studies showing that fairness-aware machine learning can have a positive effect on the reduction of discriminatory outcomes with minimal degradation (Ashokan & Haas, 2021). Further post-processing techniques such as threshold adjustment and applicant re-ranking can be leveraged to improve the fairness of hiring decisions. These strategies satisfy fairness restrictions such as demographic parity and equal opportunity (Feldman et al., 2015). Thus, the proposed method proposes a multi-layer biased reduction mechanism which is robust for both the fairness and the dependability. The system shows a good trade-off between accuracy, precision and recall and improves fairness metrics in terms of performance. There is probably a little trade-off between justice and accuracy, but the overall performance of the model is satisfactory which shows the feasibility of solving fairness-aware AI in recruitment process.

4.2 Discussion

The findings of this article highlight the importance of incorporating fairness at each stage of the AI lifecycle in recruitment tools. The proposed model differs from the traditional hiring models where fairness is a secondary objective and incorporates ethical considerations at every step of the process, from the handling of data to the final selection. This view is in line with a major limitation in the current literature that treats strategies to minimize bias independently (Mehrabi et al., 2021). The main conclusion of this study that justice and efficiency is not a conflict. Previous work has shown that there is a large trade-off between the predicted accuracy and fairness, but more recent work suggests that we should expect to achieve an acceptable balance through carefully designed algorithms (Verma & Rubin, 2018). The proposed approach supports such a view, by producing comparable results in terms of accuracy but improved fairness measures. It also points to transparency and interpretability as key features that build trust in AI-driven hiring. The incorporation of explainable AI (XAI) approaches facilitates HR management to comprehend and query the algorithmic outcomes, thus, enhancing responsibility and diminishing concerns around the application of AI (Rudin, 2019). Such transparency is even more important in high stakes decisions such as recruiting for which "one size fits all" techniques are problematic and fairness and ethics concerns are more pertinent. It also means that human-centered supervision of AI-based recruiting techniques is necessary. AI makes it more efficient and consistent, but it needs human judgment. Our proposed RH frameworks have human in the loop decision making, where HR managers can review and validate the algorithm's recommendations. This hybrid approach is in line with the conception of augmentation where AI is meant to augment humans and not replace human decision

making (Tambe et al., 2019). The paper acknowledges the challenge of creating fair hiring algorithms but does not solve any of the challenges in its contribution. One difficulty is the lack of unbiased data of the right quality. AI models are trained on historical data, and the outcome can be influenced by biases in the data.

Moreover, fairness is still hard to define and evaluate, since different responses can be obtained by using different fairness measures (Binns, 2018). Moreover, the labor market and organizations are dynamic and, therefore, AI systems should be regularly checked and maintained. Biases may change over time, so it may be necessary to perform regular analysis and update algorithms. This necessity further highlights the need for a life cycle view of AI governance where fairness is constantly assessed and improved.

4.3 Implications

The consequences of the investigation have considerable theoretical and practical implications. The paper theoretically contributes to the AI in HRM literature by proposing a multi-dimensional fairness framework (technical, ethical and organizational) to comprehend fairness in AI-based recruitment. This study advances the previous literature by showing how a collection of debiasing procedures can be methodically combined to improve fairness in selection. In practice, the proposed algorithm provides a methodology for HR managers to use a fair AI system. With this approach, firms may create recruitment solutions that enable them to hire for diversity, equity and inclusion without losing efficiency or effectiveness. The framework also supports compliance with evolving legislation and ethical expectations around AI use in hiring.

5. Limitations and Future Research Directions

5.1 Limitations

This research follows a conceptual design science path and therefore does not empirically test the proposed model with organizational recruitment data. The approach is based on a rich body of existing literature in algorithmic fairness, bias reduction, explainable AI and HRM governance, but its practical utility in real-life recruitment contexts has yet to be shown. Therefore, the takeaways presented herein should be understood as a theoretically informed framework and not a fully validated operational model. Future research may expand pilot implementations, simulation-based experiments, case studies, and large recruiting datasets to explore the potential of the framework to improve fairness, transparency, and positive hiring outcomes across a range of organizational contexts.

5.2 Future Research Directions

One particularly interesting direction for future work is to test the proposed fair hiring algorithm on real-world data sets. Further work may include exploring more advanced AI approaches (e.g., deep learning, generative AI) towards improving fairness and accuracy. Moreover, scientists may assess the impact of cultural, legal and organizational factors on fairness in AI hiring systems. Longitudinal studies also have the advantage of being able to determine whether bias reduction techniques are persistent over time.

6. Conclusion

This study highlights the growing significance of ethical AI in the recruitment industry and proposes a fair hiring algorithm with bias mitigation and fairness considerations. The findings suggest that fairness of hiring decisions can be improved without compromising efficiency by embedding fairness in the AI lifecycle. The research offers a theoretical and practical contribution by creating a conceptual framework for responsible deployment of artificial intelligence (AI) in human resource management. In summary, organizations should be transparent and accountable to ensure fair treatment and allow the continued usage of AI in today's recruitment system.

References

1. Adadi, A. and Berrada, M. (2018) Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052
2. Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management, 58*(5), 102646. https://doi.org/10.1016/j.ipm.2021.102646
3. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/
4. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 81, 149–159. https://proceedings.mlr.press/v81/binns18a.html
5. Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? Business Horizons, 63(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.0021
6. Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn.
7. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learningalgorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512
8. Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? Industrial and Organizational Psychology: Perspectives on Science and Practice, 9(3), 621–640. https://doi.org/10.1017/iop.2016.6

9. Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
10. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
11. European Commission. (2021, April 21). Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) (COM/2021/206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
12. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268). https://dl.acm.org/doi/abs/10.1145/2783258.2783311
13. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707.
14. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2019). A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\)* (pp. 329–338). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287589
15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93, 1–42. https://doi.org/10.1145/3236009
16. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://doi.org/10.48550/arXiv.1610.02413
17. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
19. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.https://doi.org/10.2307/j.ctt1pwt9w5
20. Ntoutsi, E., Fafalios, P., & Gadiraju, U. (2020). Bias in datadriven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), e1356. https://doi.org/10.1002/widm.1356
21. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

22. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372873
23. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. The Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072
24. Rieke, A. & Bogen, M. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn. https://www.upturn.org/reports/2018/hiring-algorithms/
25. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x
26. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. California Management Review, 61, 15-42. https://doi.org/10.1177/0008125619867910
27. Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. Strategic HR Review, 17(5), 255–258. https://doi.org/10.1108/SHR-07-2018-0051
28. Verma, S., & Rubin, J. (2018). Fairness definitions explained. In FairWare '18: Proceedings of the International Workshop on Software Fairness (pp. 1–7). ACM. https://doi.org/10.1145/3194770.319477
29. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 842–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf
30. Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171–1180). Association for Computing Machinery. https://doi.org/10.1145/3038912.3052660